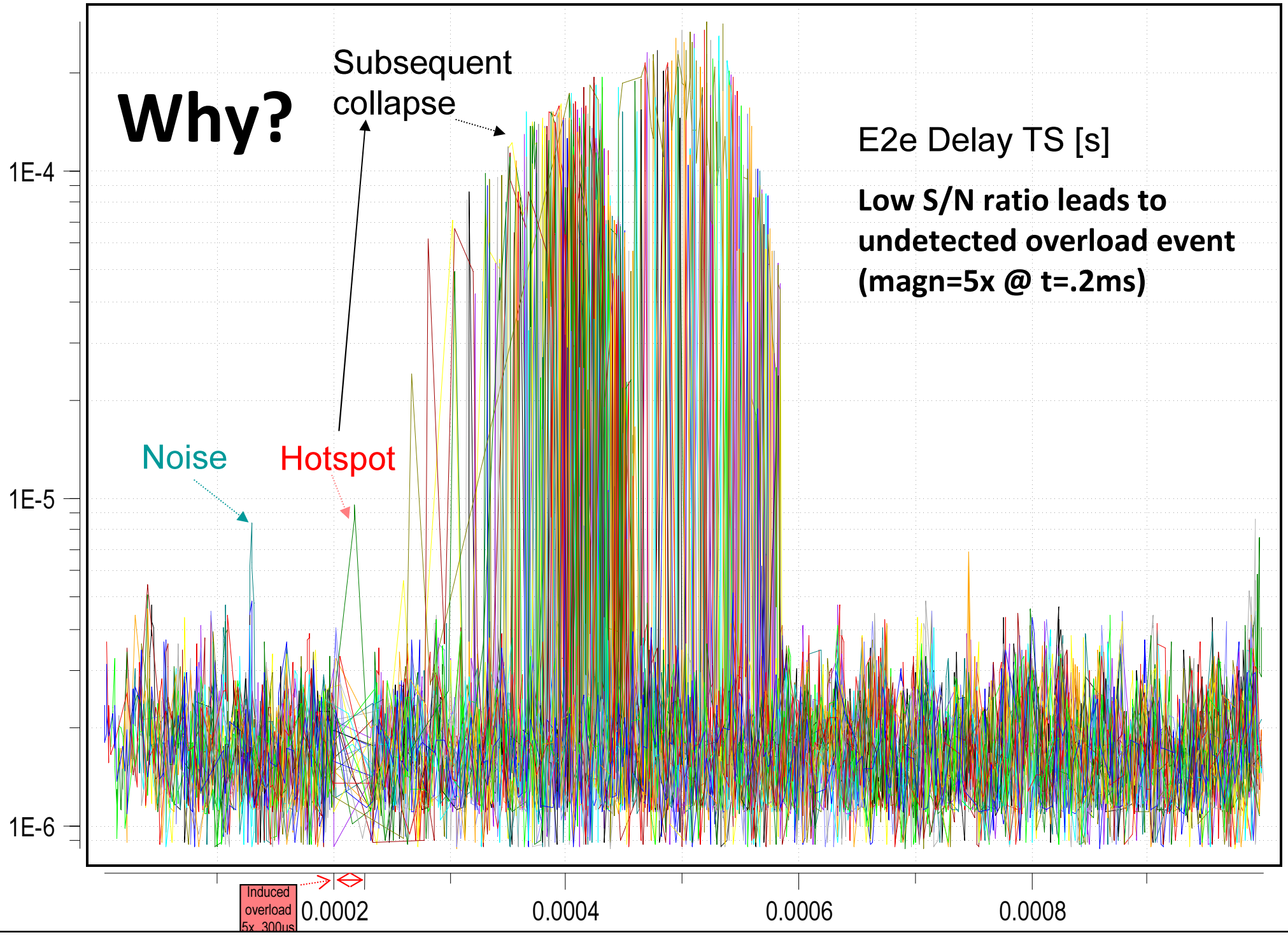
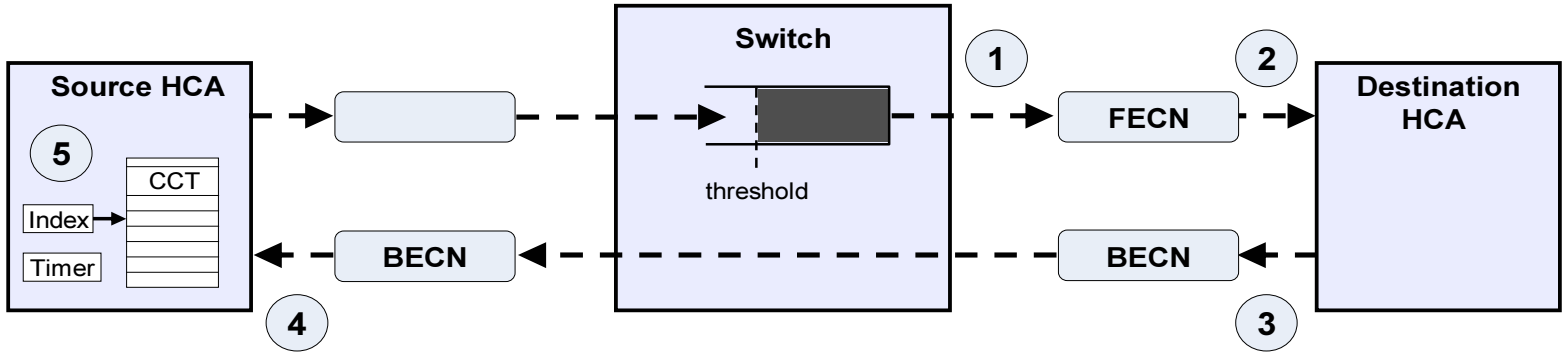
Control the (1) Rate, (2) Route, (3) Globally/SDN, (4) Locally… ?



Control in the t-dimension → Rate #1: IBA CCA

- IBTA's CM (aka CCA) already shaped

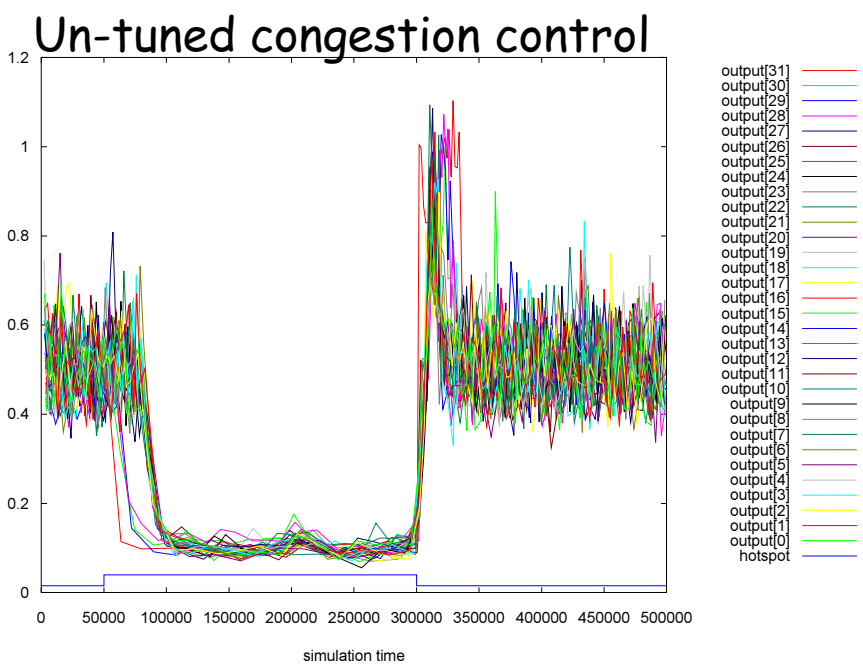


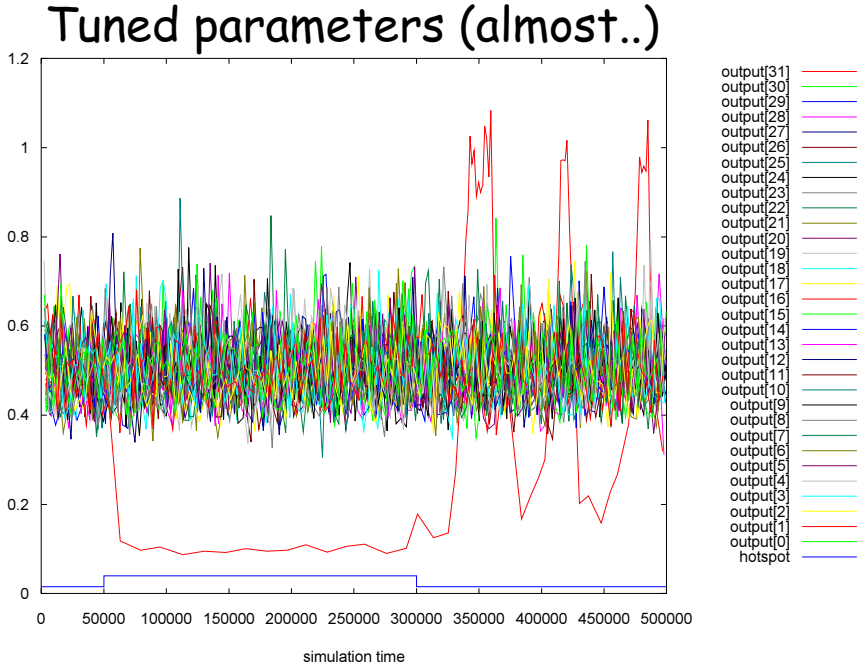
1. Load Sensor (LS): Q-occupancy;
2. Feedback (Fb): FECN; binary; single closed loop –Fb;
3. Source response function (SRF):
   1. down rate ~ FECN IA;
   2. up rate = timer based self recovery

Closely related to ECN/RED/TCP (and also DC-TCP/Sigcomm'10, and RoCEv2/Sigcomm'15)

Does CCA work? ..got a PhD to spare..? ☺

- Qualified "yes" => needs tuning
  - easy for small fabrics w/ simple traffic, hard for others...
- Param *tuning* required per (1) fabric architecture and (2) traffic
- Narrow stability: CCA sensitivity to (1,2) and params



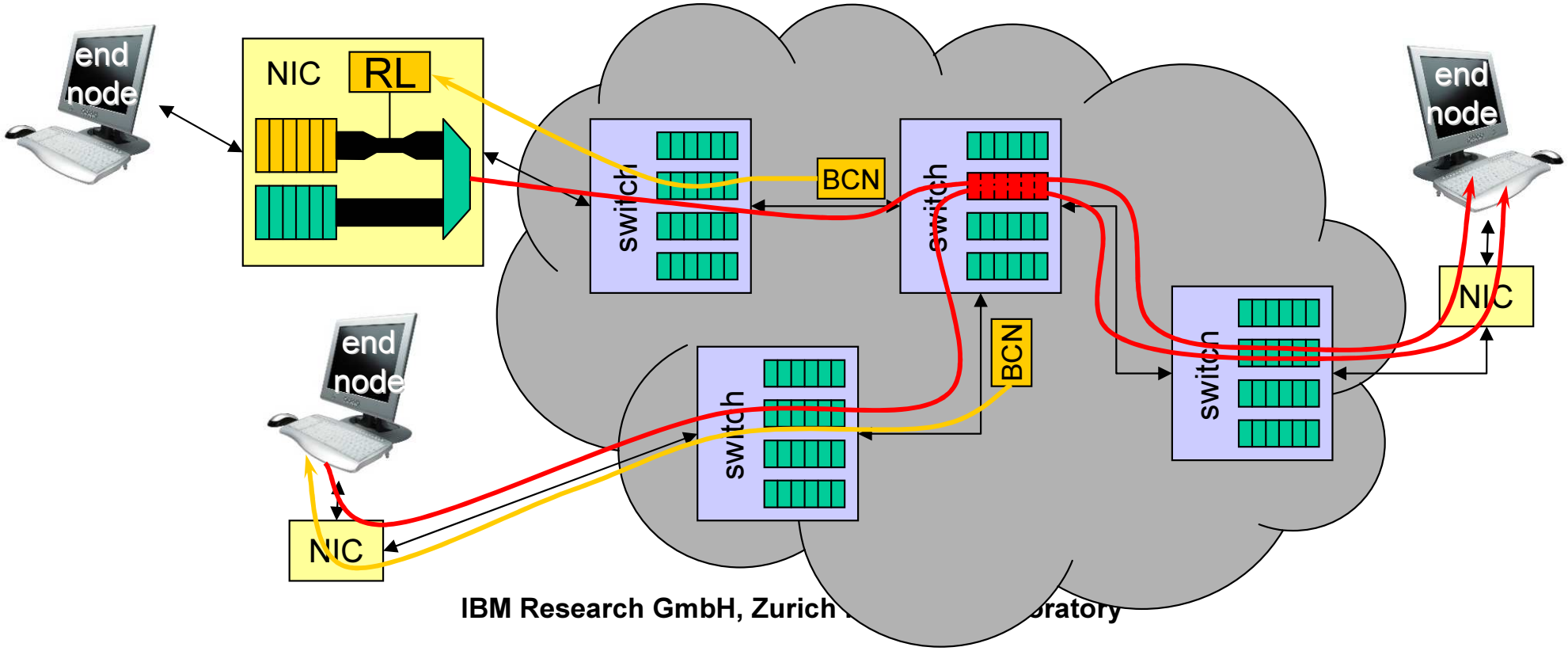


Rate #2: Ethernet QCN

1. Congestion point (CP)
   - Sampling: Q occupancy {pos,veloc} ~ 2D congestion vector
   - Derive feedback value (by applying PID and compensation, see next)

2. Feedback channel
   - Convey congestion notifications from CP **directly** to the culprit sources of "offending" traffic
   - **Multibit** Cong. Notifications contain congestion information, incl. a feedback value (copied by DC TCP)

3. Reaction point (RP)
   - Use rate limiters (RL) at the edge to shape flows causing congestion (also used by RoCEv2 et al.)
   - Adjust rates based on the multibit feedback values received from congestion points

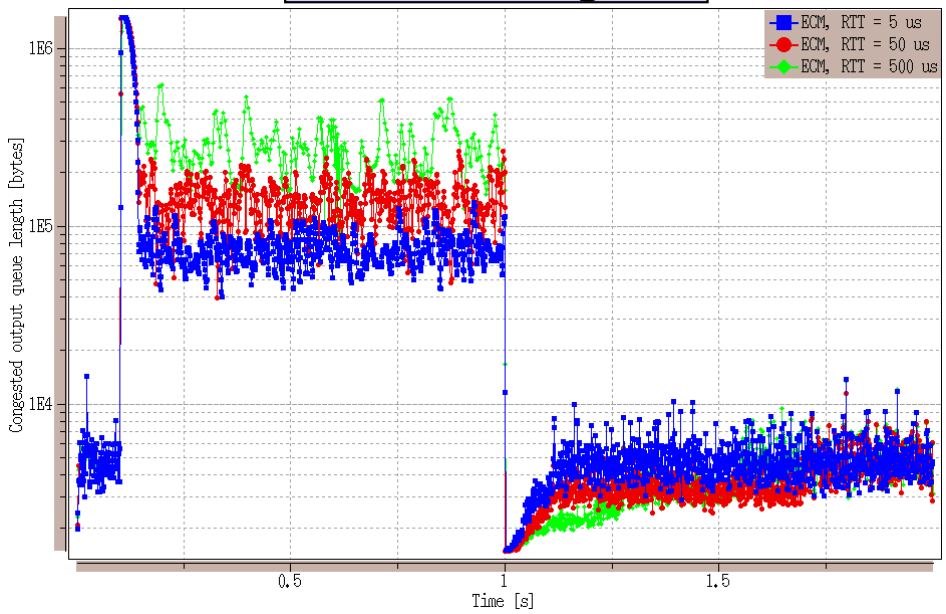


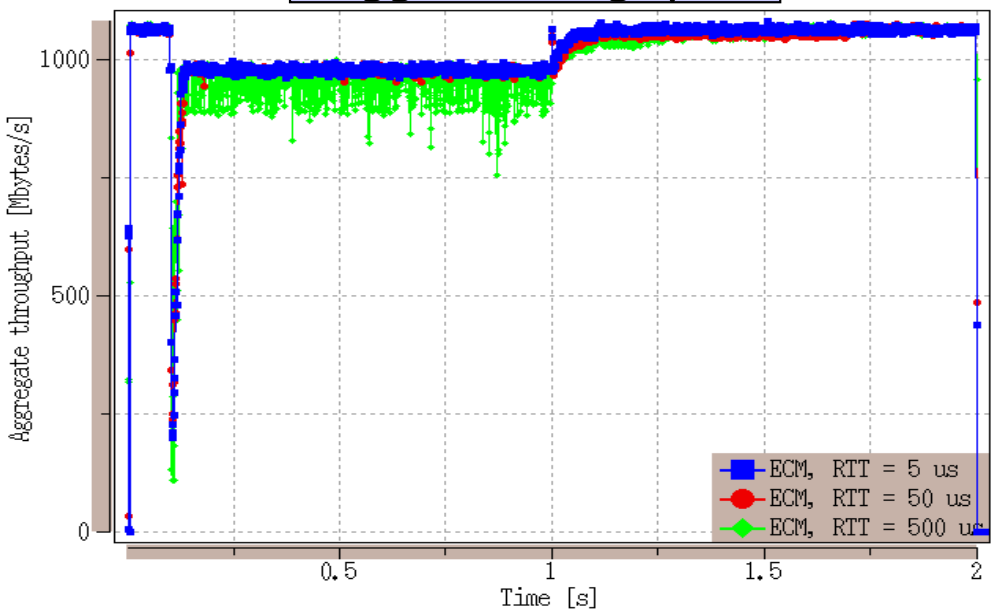
OG Hotspot Performance

Queue length

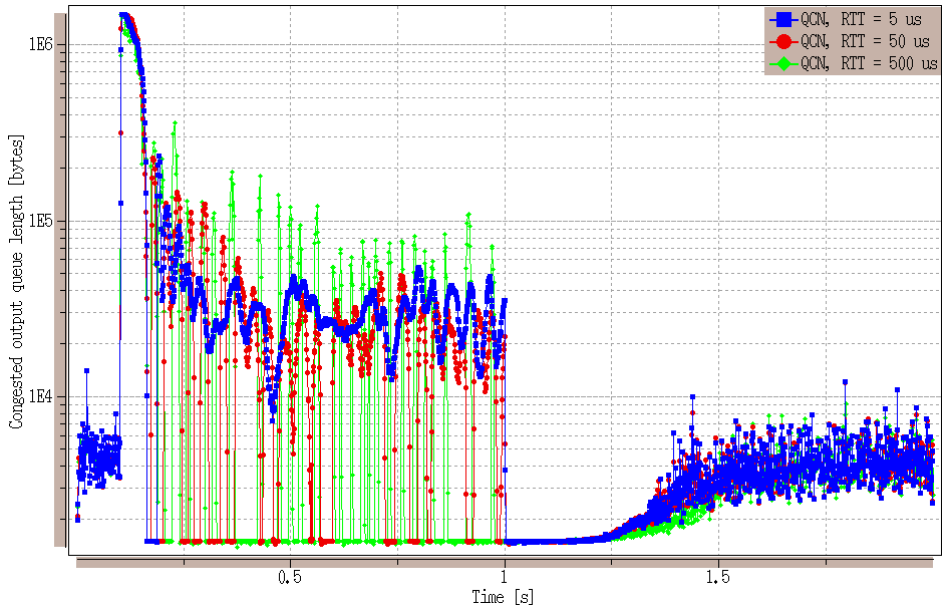
Aggr. throughput

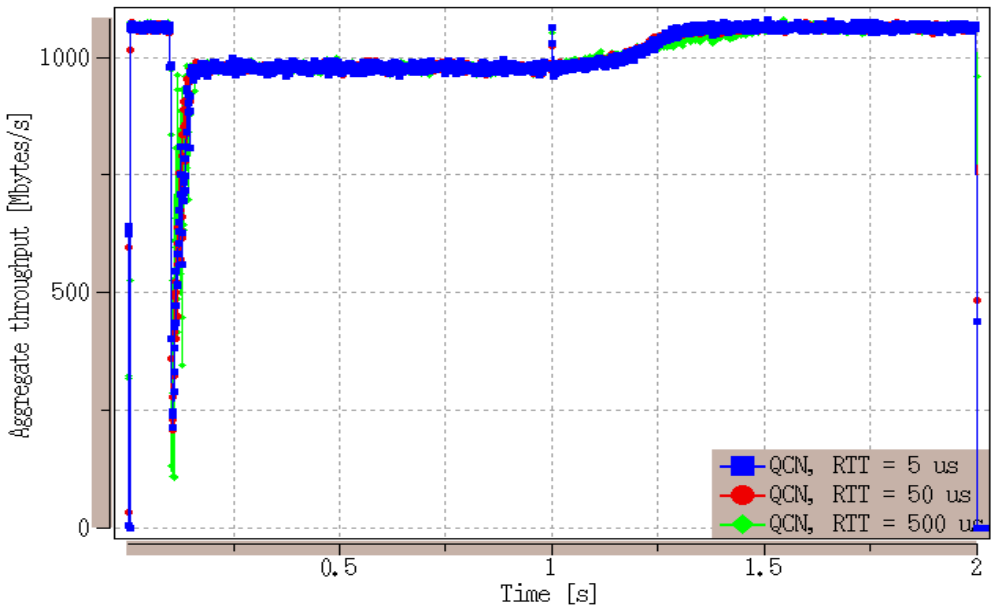
ECM





QCN





QCN's Parms: got another PhD to Spare? ☺

Parameter	Value	Unit	Parameter	Value	Unit
TCP					
buffer size	128	KB	TX delay	9.5	μs
max buffer size	256	KB	RX delay	24	μs
default RTO	10	ms	timer quanta	1	μs
min RTO	2	ms	reassembly queue	200	seg.
RTO variance	20	ms			
ECN-RED					
min thresh.	25.6	KB	W_q	0.002	
max thresh.	76.8	KB	P_{max}	0.02	
QCN					
Q_{eq}	20 or 66	KB	fast recovery thresh.	5	
W_d	2		min. rate	100	Kb/s
G_d	0.5		active incr.	5	Mb/s
CM timer	15	ms	hyperactive incr.	50	Mb/s
sample interval	150	KB	min decr. factor	0.5	
byte count limit	150	KB	extra fast recovery	enabled	
PFC					
min thresh.	80	KB	max thresh.	97	KB
Network hardware					
link speed	10	Gb/s	adapter delay	500	ns
frame size	1500	B	switch buffer size/port	100	KB
adapter buffer size	512	KB	switch delay	100	ns

I wonder why "Nobody uses 'my' congestion controls"...?

Next, how about spatial control, i.e., Routing?

Comparative Evaluation of CEE-based Adaptive Routing

Daniel Crisan, Mitch Gusat and Cyriel Minkenberg

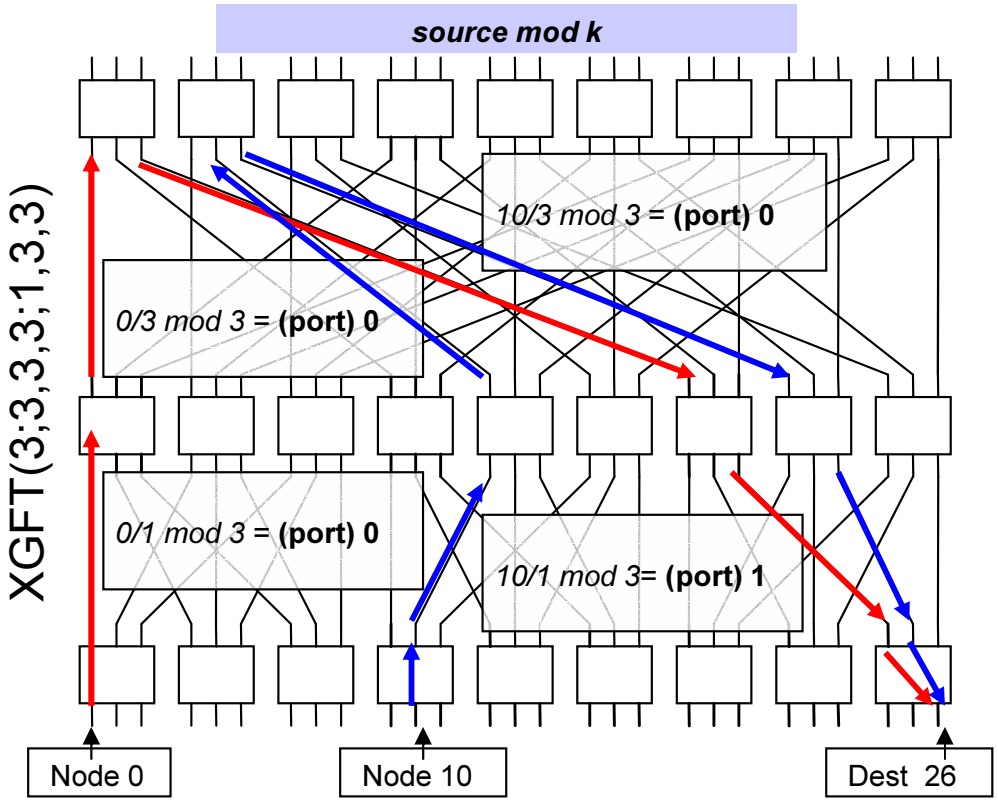
IBM Research GmbH, Zürich Research Laboratory

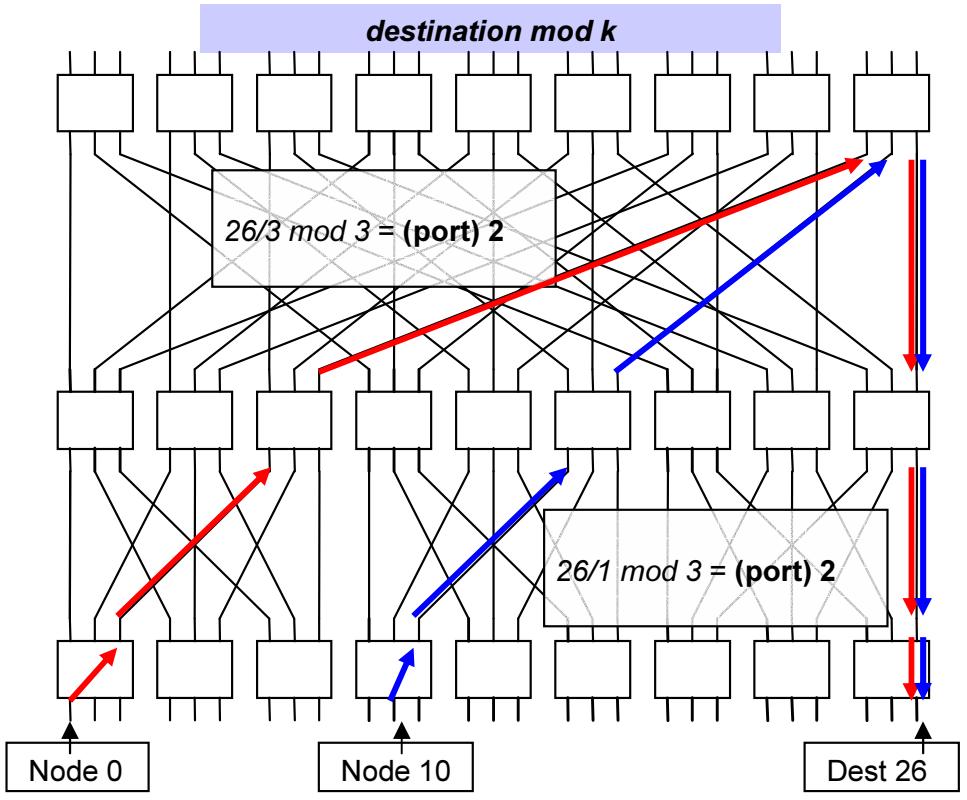
Rate or Route?
Congestion Management vs. Adaptive Routing

- CM solves congestion by reducing injection rate
  - Useful for saturation tree congestion, where many “innocent” flows suffer because of backlog of some hot flows
  - Does not exploit path diversity
  - Typical data center topologies offer high path diversity
    - Fat tree, mesh, torus
- Adaptive routing (switch AR) basic approach
  - Allow multi-path routing
  - By default route on shortest path (latency)
  - Detect downstream congestion by means of QCN
  - In case of congestion
    - First try to reroute hot flows on alternative paths
    - Only if no uncongested alternative exists, reduce send rate

Extended generalized fat tree (XGFT) topology

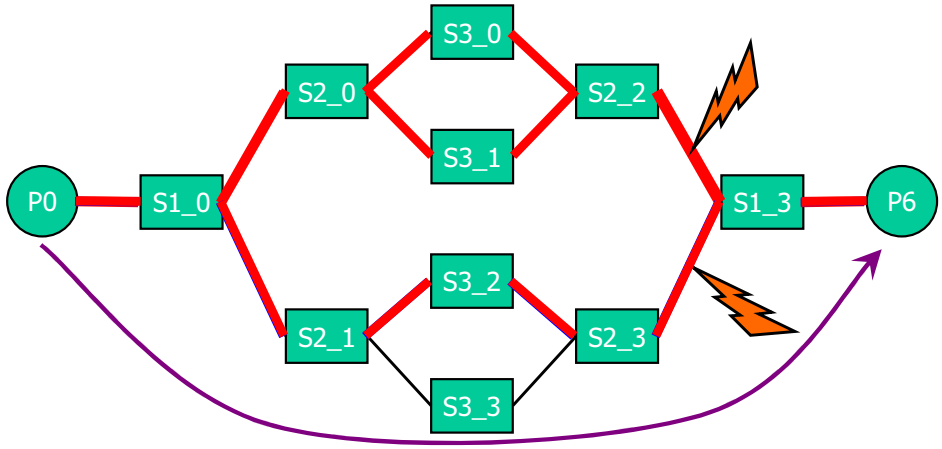
- Multi-path: one path via each top-level switch
- Self-routing
- Usual static, oblivious routing method based on label of source or destination node to select path; can lead to significant contention
- Problem of assigning paths to connections with min. number of conflicts
  - Non-oblivious offline route optimization taking into account traffic pattern

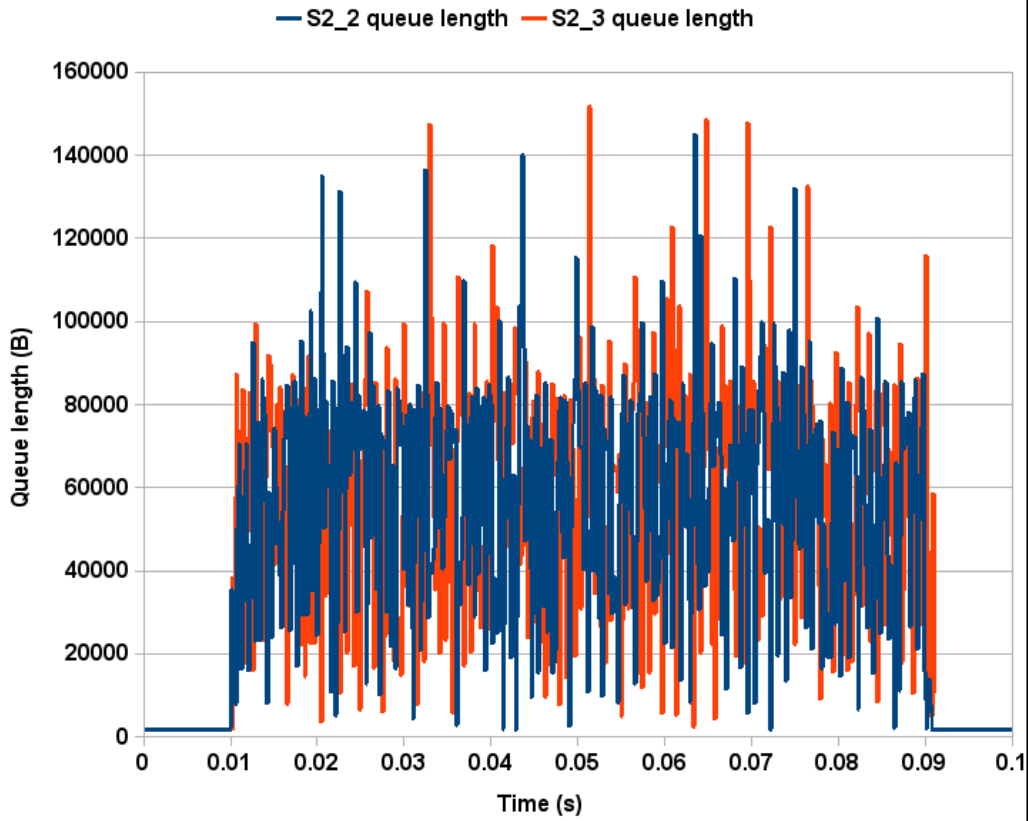




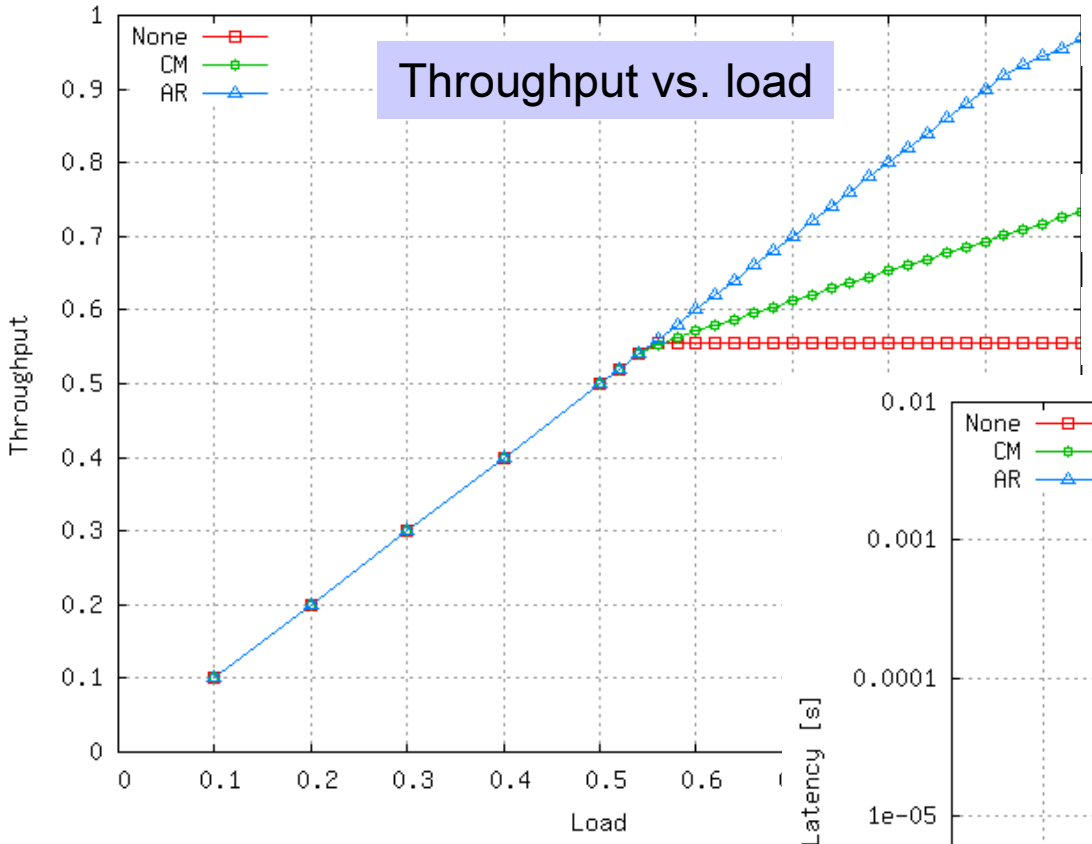
Switch Adaptive Routing

- QCN feedback provide "congestion price"
- Algorithm
  [Minkenberg&Gusat'09]
  - switches snoop the CNs
  - based on feedback – steer the traffic
- Advantages
  - Congestion avoidance
  - Use of alternative paths
- Oscillations possible
- Routing controlled by switches





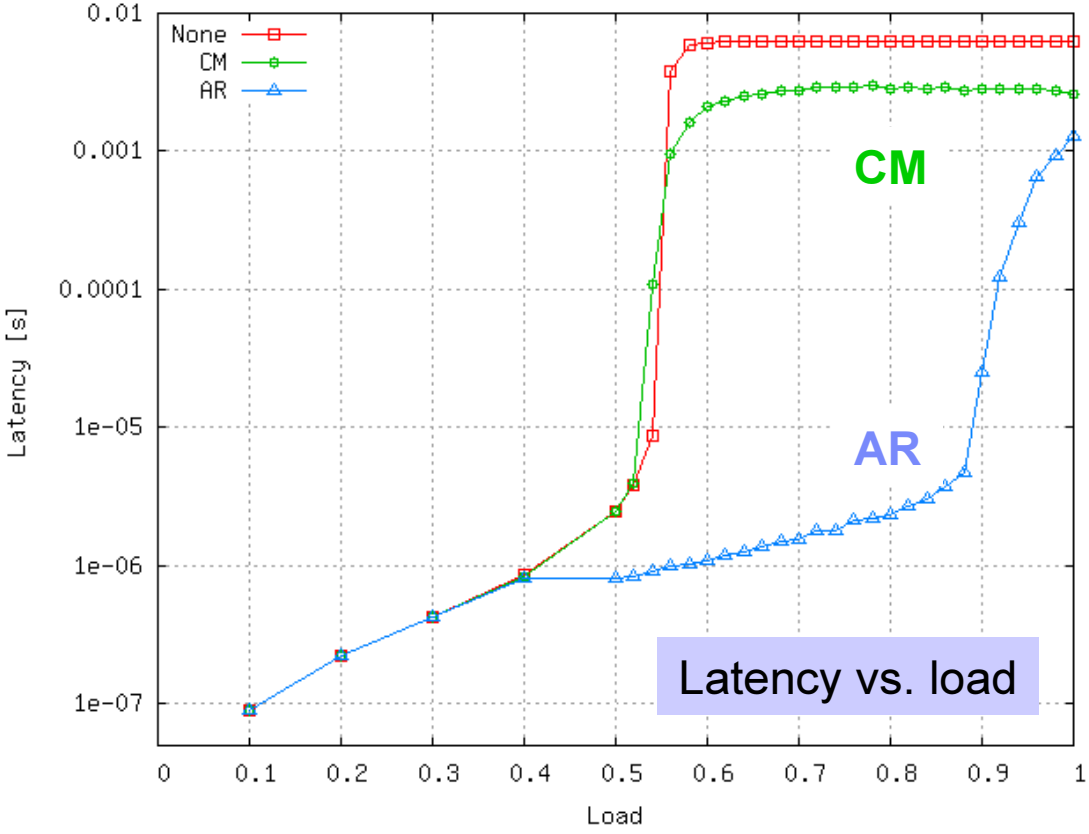
Rate/CM vs. Route/AR: Bernoulli Traffic Simulation



AR

CM

None



None

CM

AR

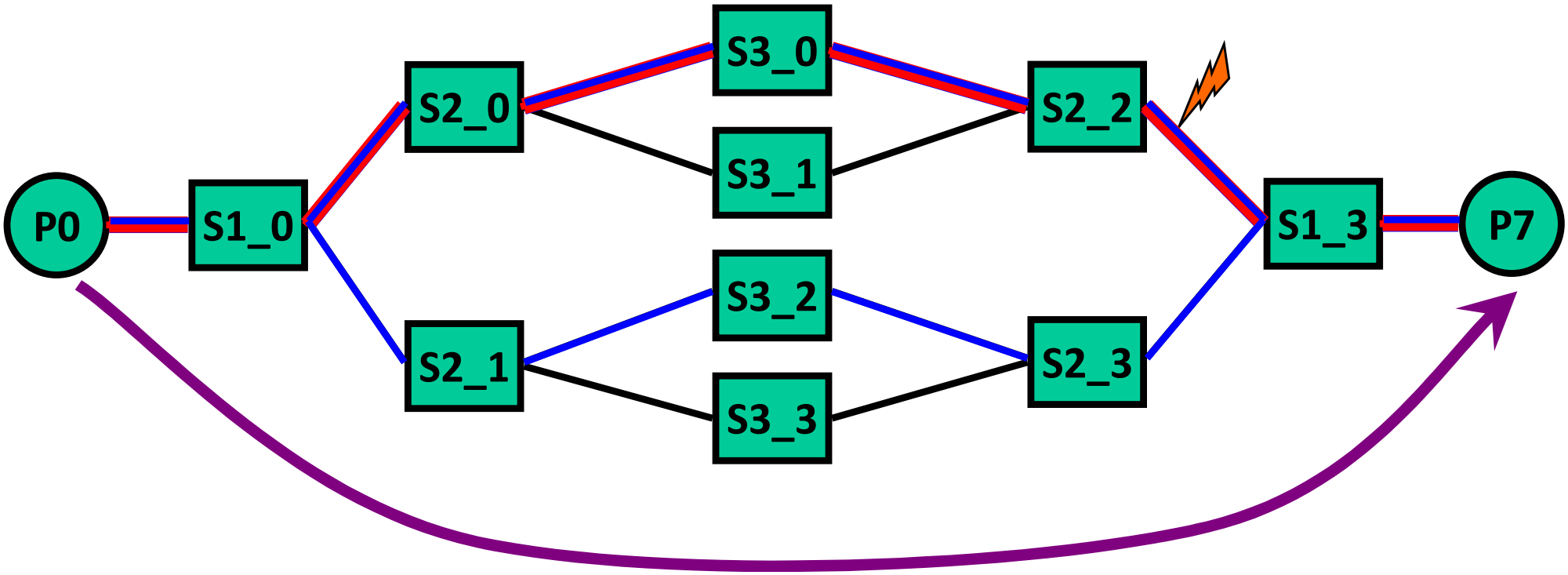
Source AR: R^3C^2 Concept

Take advantage of CNMs at the source for adaptive load-balancing

- **Congestion Point** issues CNMs
  - Where is the hotspot?
  - How severe is the hotspot?

- **Source** receives the CNMs
  - Identifies the most severe hotspots
  - Reroutes traffic around the hotspots
  - Splits flows and rate-limits subflows

R^3C^2 Algorithm



- No overload: Deterministic single path
- Congestion: Activate additional paths
- Path activation: avoid hotspots
- Use RL along each path

Evaluation Methodology

- Venus + Dimemas simulator

- Traffic
  - Synthetic: permutations + hotspot
  - HPC Traces:
    - NAS: BT, CG, FT, IS, MG
    - WRF, NAMD, Liso, Airbus

- Model parameters
  - 10Gbps CEE with MTU = 1500B
  - QCN and PFC: 802 DCB settings

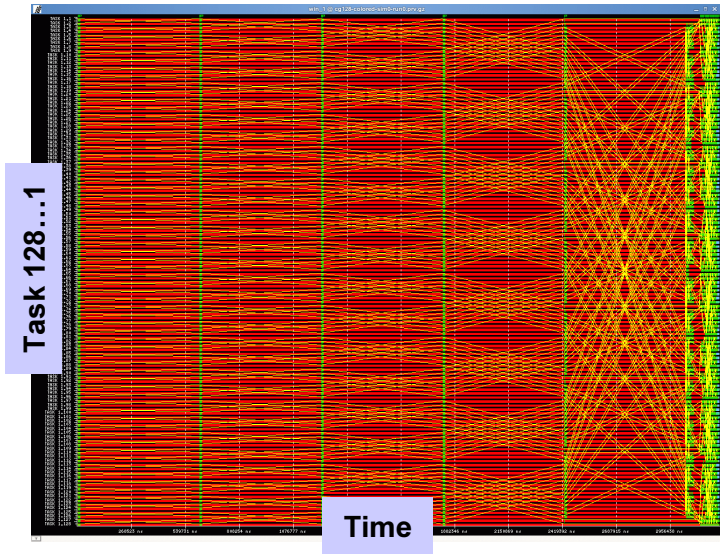
- Topology: 2-ary n-tree

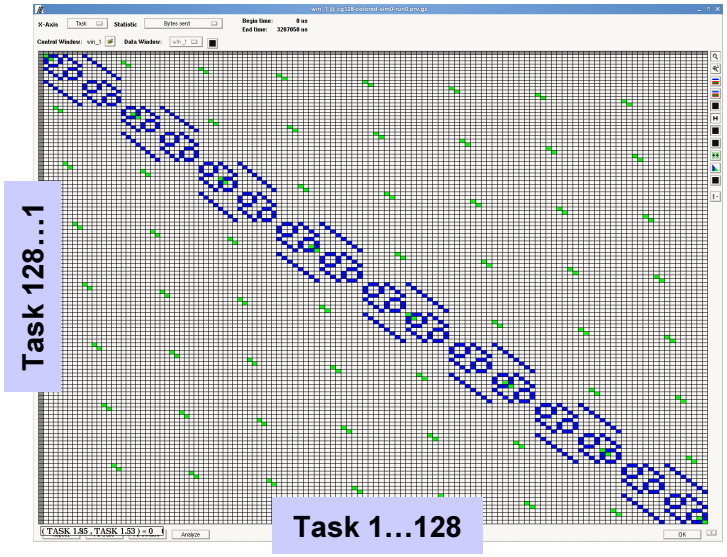
CG and FT communication patterns

Communication pattern

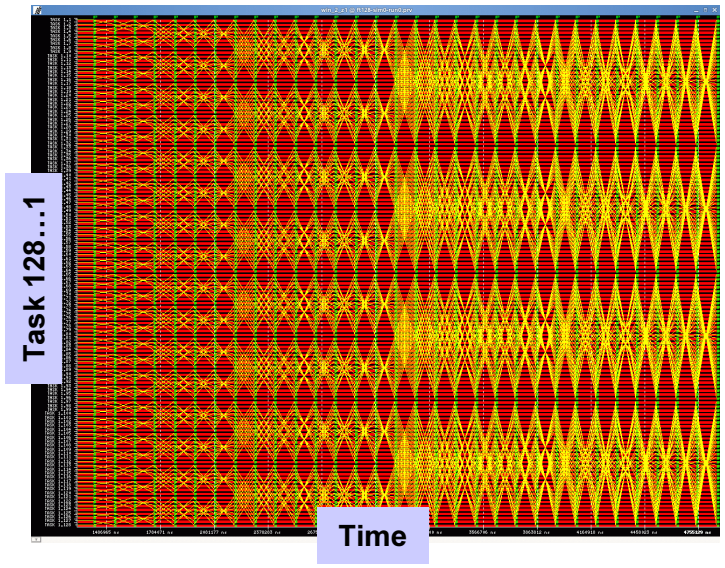
Traffic volume per node pair

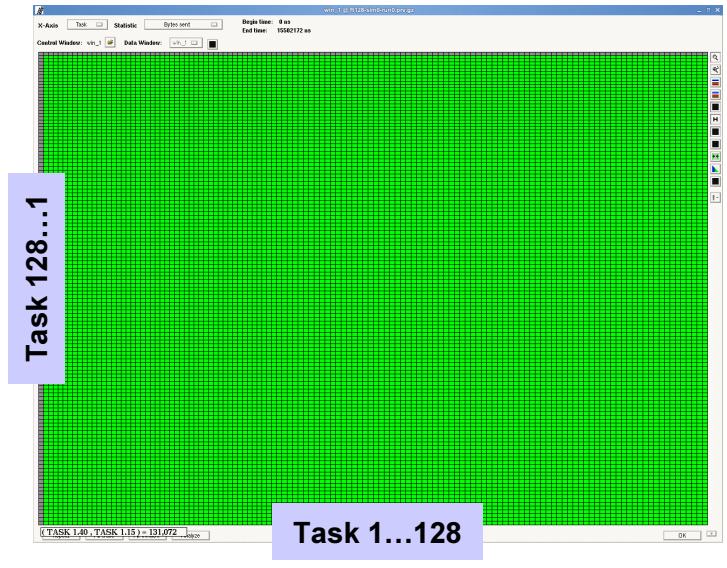
Conjugate Gradient



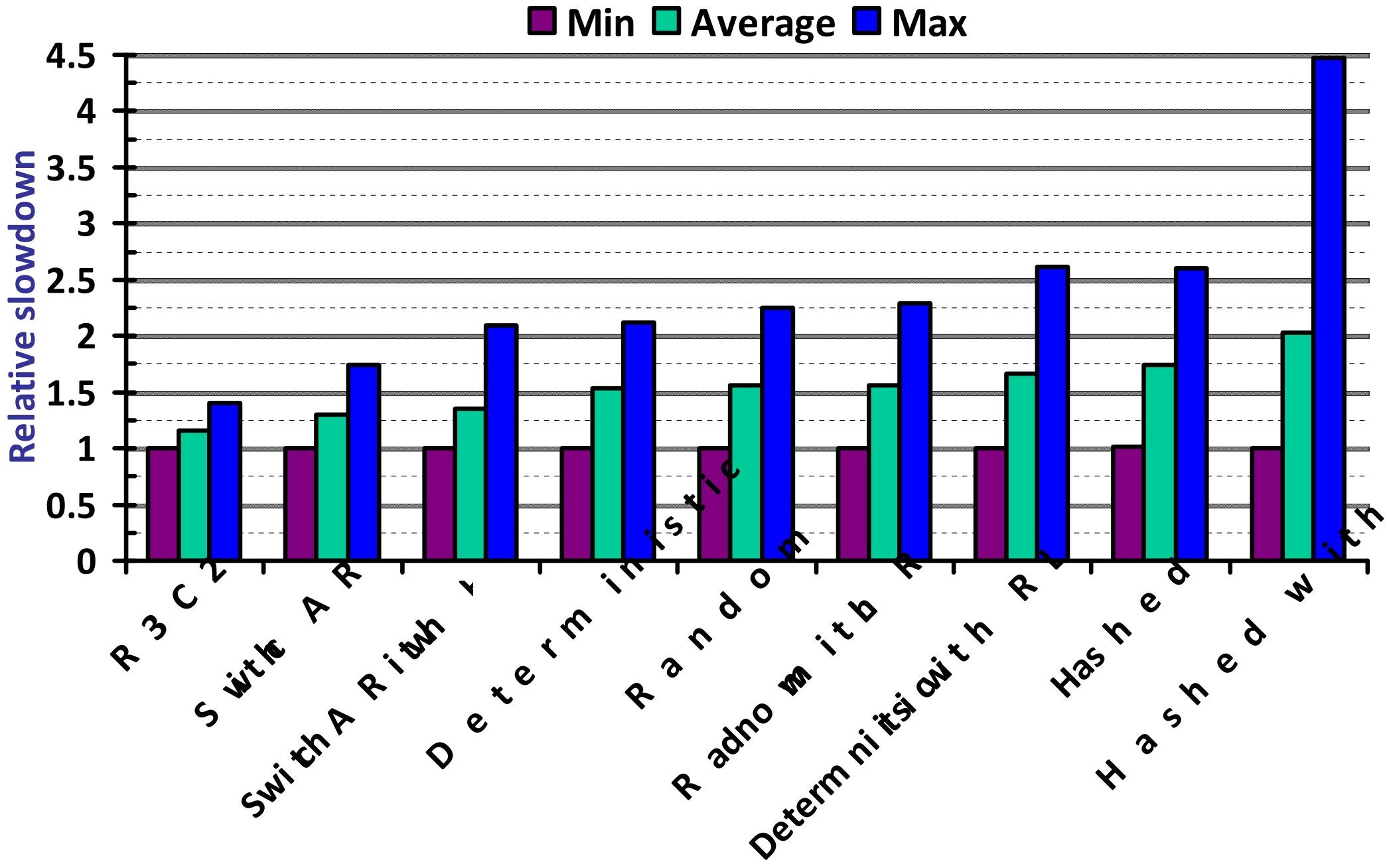


Fast Fourier Transform





HPC Traces: Hotspot



Rate or Route Control? Local or Global?

- Many topos offer abundant multipaths
  - Load balancing and reliability options

- Best routing: a qualified answer…
  1. D-mod-k deterministic: simple + no OOO delivery
  2. Random (-OOO) and hash: win under ideal DCN conditions, single prio, no failures or local overloads, w/ 'easy' traffic
  3. Adaptive (-OOO): best trade-off under realistic DCN scenarios… Performance benefits:

     80% over Deterministic

     40% over Random
- Rate or route ➔ Dual Route & Rate control
  - Improved stability and performance
- Open: ordering and additional cost vs. hashing